Max-Planck-Institut für demografische Forschung

Max Planck Institute for Demographic Research

Konrad-Zuse-Strasse 1 ▪ D-18057 Rostock ▪ Germany ▪ Tel +49 (0) 3 81 20 81 - 0 ▪ Fax +49 (0) 3 81 20 81 - 202 ▪ www.demogr.mpg.de

MPIDR Working Paper WP 2018-001 I January 2018

New methods for estimating detailed fertility schedules from abridged data

Pavel Grigoriev I grigoriev@demogr.mpg.de
Anatoli I. Michalski
Vasily P. Gorlishchev
Dmitri A. Jdanov
Vladimir M. Shkolnikov

Human Fertility Database Research Report
HFD RR-2018-001

New methods for estimating detailed fertility schedules from abridged data

Pavel Grigoriev[1], Anatoli I. Michalski[2], Vasily P. Gorlishchev[2], Dmitri A. Jdanov[1,3], Vladimir M. Shkolnikov[1,3]

[1] Max Planck Institute for Demographic Research, Rostock, Germany
[2] V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia
[3] National Research University Higher School of Economics, Moscow, Russia

ABSTRACT

Background and Aim Occasionally, there is a need to split aggregated fertility data into a fine grid of ages. For this purpose, several disaggregation methods have been developed. Yet these methods have some limitations. We seek to identify a method that satisfies the following criteria: 1) *shape* – the estimated fertility curves should be plausible and smooth; 2) *fit* – the predicted values should closely trace the observed values; 3) *non-negativity* – only positive values should be returned; 4) *balance* – the estimated five-year age group totals should match the input data; and in case of birth order data 5) *parity* – the balance by parity has to be maintained. To our knowledge, none of the existing methods fully meets the first four criteria. Moreover, no attempt has been made to extend the restrictions to criterion (5). To address the disadvantages of the existing methods, we introduce two alternative approaches for splitting abridged fertility data: namely, the *quadratic optimization* (QO) method and the *neural network* (NN) method. **Data and Methods** We mainly rely on high-quality fertility data from the Human Fertility Database (HFD), Additionally, we use a large and heterogeneous dataset from the Human Fertility Collection (HFC). The performance of the proposed methods is evaluated both visually (by examining of the obtained fertility schedules), and statistically using several metrics of fit. The QO and NN methods are tested against the current HFD splitting protocol (HFD method) and the calibrated spline (CS) method. **Results** The results of thorough testing suggest that both methods perform well. The main advantage – and a distinguishing feature – of the QO approach is that it meets all of the requirements listed above. However, it does not provide a *fit* as good as that of the NN and CS methods. In addition, when it is applied to birth order data, it can sometimes produce implausible shapes for parity 1. To account for such cases, we have developed individual solutions, which can easily be adapted to account for other cases that might occur. While the NN method does not satisfy the *balance* and *parity* criteria, it returns better results in terms of *fit* than the other methods. **Conclusions** The QO method satisfies the needs of large databases such as the HFD and the HFC. While this method has very strict requirements, it returns plausible fertility estimates regardless of the nature of the input data. The NN method appears to be a suitable alternative for use in individual cases in which the priority is given to the *fit* criterion.

1. Introduction

The problem of splitting aggregated fertility data into single years of age is often encountered by demographers. To address this issue, several disaggregation methods have been developed (McNeil et. 1977; Smith, Hyndman, Wood, 2004; Liu, et al. 2011; Schmertmann, 2012; Jasilioniene et al. 2012). Using a sample of HFD countries, Liu et al. (2011) tested 10 different methods that derive age-specific fertility rates from abridged data, and concluded that the modified Beers method (de Beer, 2011) provided the best fit. Using the HFD and the US Census International Database, Schmertmann (2012) compared the performance of the calibrated spline (CS) with that of the Beers and HFD methods. The analysis showed that while the three methods performed very well, the CS method provided the best fit. In the overall ranking, the CS method placed first, the HFD method placed second, and the Beers method placed third (Schmertmann, 2012).

The disaggregation problem primarily occurs in relation to historical data and data from developing countries that lack functioning systems of vital registration. Splitting is often required for the purposes of harmonizing the data so that they are comparable across time and countries. This issue is particularly relevant for the maintenance of large international databases, such as the Human Fertility Database[1] and the Human Fertility Collection[2]. At the moment, the HFD has its own splitting protocol, the HFD method (Jasilioniene et al. 2012); while the HFC uses the CS estimator to disaggregate age-specific fertility rates (Grigorieva et al. 2015). According to Schmertmann (2012), these two methods appear to produce the best results. The effectiveness of these methods has also been confirmed by extensive experiments with real data from the HFD and the HFC (Grigoriev and Jdanov, 2015). There are, however, several important differences between the HFC and the HFD that affect the choice of estimation strategy:

1) The degree of heterogeneity of the input data. The HFD contains much more homogenous fertility data than the HFC, which gathers all available fertility data across the globe, and has low data quality requirements.

2) The target measure to be estimated. In the HFD it is the absolute number of births, while in the HFC – age-specific fertility rates.

[1] http://www.humanfertility.org
[2] http://www.fertilitydata.org

3) The HFD provides high-quality data that allows for high-quality research. The interpolation should not be transferred into smoothing, which might remove real effects.

4) The original data included in the HFC is likely to be noisy and erroneous, especially for countries without a functioning system of vital statistics. In such cases, smoothing is a good solution.

While both the HFD and the CS methods meet the basic requirements of the HFD and the HFC, they are not free of limitations (We discuss these limitations in more detail in section 2). Moreover, despite the differences between the HFD and the HFC, it is more reasonable to rely on a universe splitting protocol in both databases. All of these considerations motivated us to develop alternative methods that could be universally applied to both the high-quality HFD data and the heterogeneous and noisy HFC data. As an alternative to the current HFD method, we seek to identify a method that could simultaneously satisfy the following criteria:

1) *Shape* - The estimated fertility curves should be plausible and smooth.
2) *Fit* – The predicted values should closely trace the observed values.
3) *Non-negativity* – Only positive predicted values should be returned.
4) *Balance* – The estimated five-year age group totals should match the input data.
5) *Parity* – The balance by parity has to be maintained after splitting.

Generally, the HFD method meets criteria (1)–(4), but may not always satisfy criterion (1) because of a mathematical feature of the spline function used for this method. The HFD method relies on the Hermite spline, which is monotonic, and thus meets criterion (3). However, unlike other polynomial splines, it has a discontinuous second derivative, which might result in sudden twists in the estimated fertility curves. The CS method satisfies criteria (1)-(3) but does not satisfy criterion (4), which is crucial for the requirements of the HFD. In addition, the CS method is rather complex. This lack of suitable methods motivated us to search for alternatives. To our knowledge, none of the existing methods fully meets criteria (1)-(4). Moreover, no attempt has been made to extend the restrictions to meet criterion (5).

In this paper, we address the disadvantages of the existing methods by introducing two alternative approaches for splitting abridged fertility data: namely, the *quadratic optimization* (QO) method and the *neural network* (NN) method. To assess the performance of the proposed methods, we relied on high-quality detailed fertility data from the HFD and a large sample from the HFC. These data, along with R scripts containing the QO and NN functions and various examples of their usage, are provided in the MPIDR technical reports (see Michalski, Grigoriev, Gorlischev, 2018 and Gorlischev, Grigoriev, Michalski, 2018).

2. Limitations of the HFD and CS methods

2.1. HFD Method

The HFD splitting procedure is based on the interpolation of the cumulative rates, which follows the method proposed by McNeil et al. (1977). The HFD algorithm described in the HFD Methods Protocol (see Jasilione et al. 2012, pp.30-31) consists of the following steps:

1) Calculating cumulative fertility rate $F(x)$ from age-specific fertility rates $f(x)$;
2) Calculating logits of cumulative fertility rate

$$Y(x) = LOGIT\left[\frac{F(x)}{F(x_{\max})}\right] = \log\left[\frac{F(x)}{F(x_{\max}) - F(x)}\right];$$

3) Setting $Y(x_{\min}) = -20$ and $Y(x_{\max}) = 12$ for the two data points (extremes) where the logarithm is not defined;
4) Estimating $\hat{Y}(x)$, a continuous version of $Y(x)$ using Hermite cubic spline interpolation (function 'interp1', method='pchip', R library 'signal');
5) Estimating $\hat{F}(x)$, a continuous version of $F(x)$, using inverse logit transformation:

$$\hat{F}(x) = \left[\frac{e^{\hat{Y}(x)}}{1 + e^{\hat{Y}(x)}}\right] \times F(x_{\max})$$

6) And, finally, obtaining single-year rates ${}_1f_x = F(x) = F(x+1) - F(x)$.

Figure 2.1 visualizes the main steps of this procedure. The major disadvantage of the technique described above is point (3), in which undefined logarithms at the extremes (*–Inf* and *+Inf*) have to be replaced by arbitrary values. These values – which we hereafter refer to as *LO* and *HI,* respectively – are currently set to –20 for the lower limit and +12 for the upper limit of the distribution. The testing on the HFD data has shown that in most cases, Hermite spline interpolation with these values produces satisfactory results. In some cases, however, the HFD method fails to generate plausible fertility curves. Figure 2.2 depicts such an (hypothetical) example: the ASFR estimates below age 30 appear implausible, and they are very far from the observed values.

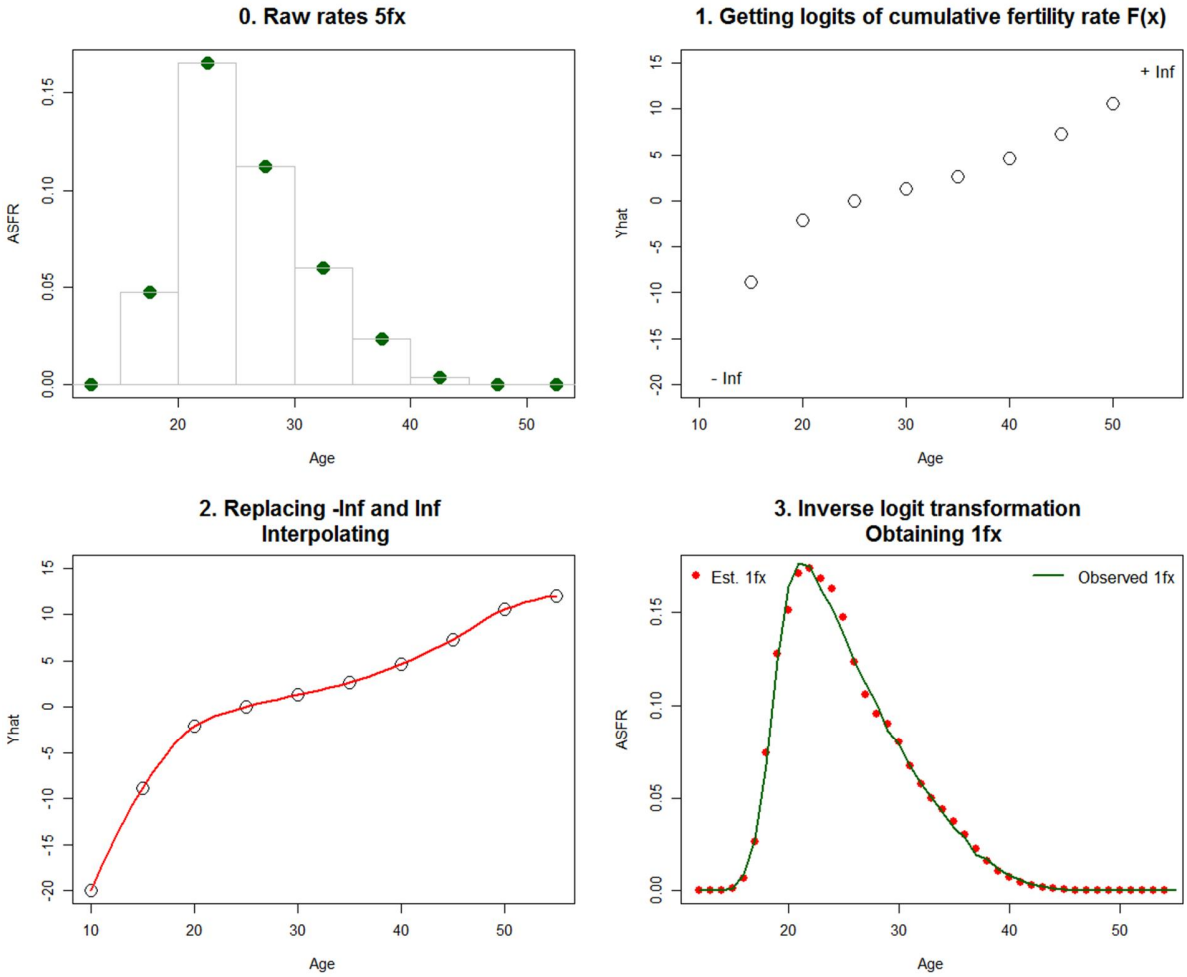


Figure 2.1

Estimating single-year age-specific fertility rates ($_1f_x$) on the basis of 5-year ($_5f_x$) age-specific fertility rates; the HFD method, Russia, 1985

Source: Human Fertility Database

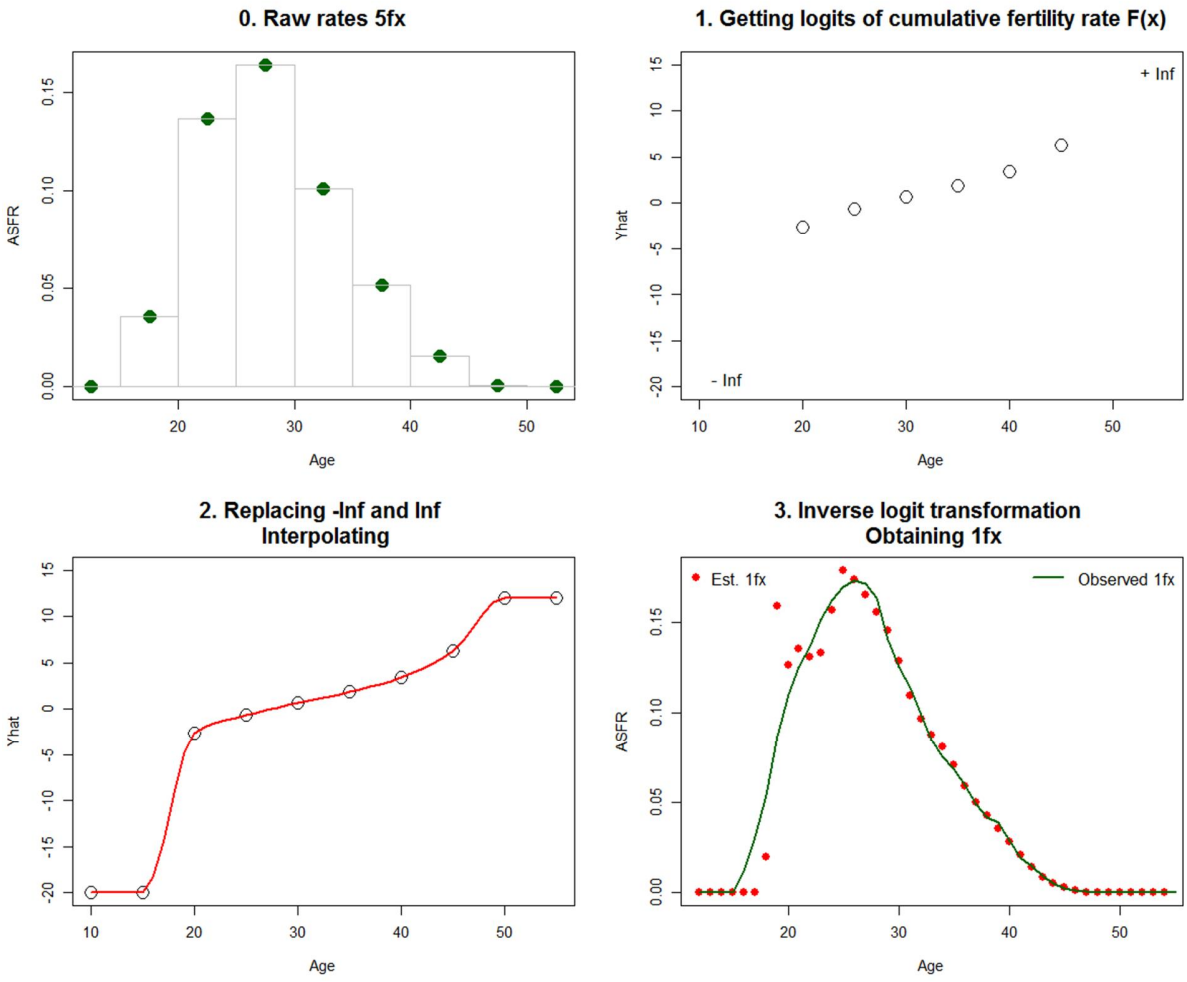


Figure 2.2

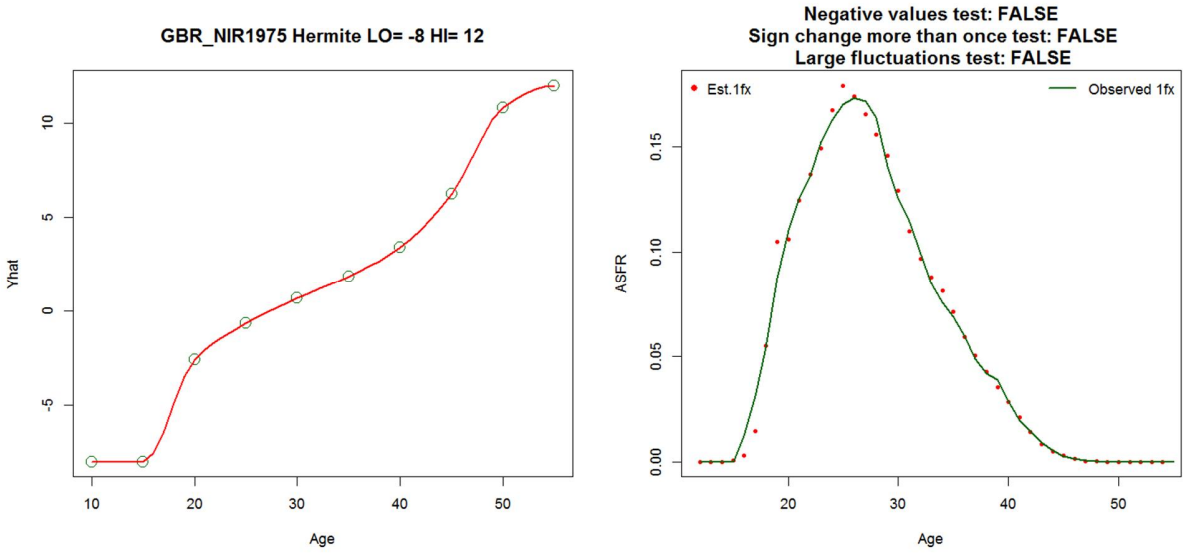
Estimating single-year age-specific fertility rates ($_1f_x$) on the basis of 5-year ($_5f_x$) age-specific fertility rates; the HFD method, Northern Ireland, 1975

Source: As for Figure 2.1

At first glance, the problem appears to be related to the mathematical properties of the Hermite spline. To our knowledge, at the moment of its adaptation for the HFD computational routine, the Hermite spline was the only method implemented in R that quarantined the non-negativity of the interpolated data. However, unlike other polynomial splines, it does not guarantee the continuous second derivative. The second derivative of interpolated cumulative fertility curve $F''(x)$ is discontinuous by definition; as is its first derivative, ASFR fertility schedule $f(x) = F'(x)$ The discontinuity of the derivatives might produce undesirable results, such as kinks and abrupt slope changes in the fertility curve. The other interpolations methods do not have this property,

but they can return negative values. Nevertheless, when the same spline function is applied but just the LO value is changed from –20 to –8, the results improve radically (Panel A of Figure 2.3).

A. Calibrating LO and HI values



B. Assigning a small artificial value (phantom birth) to the first group

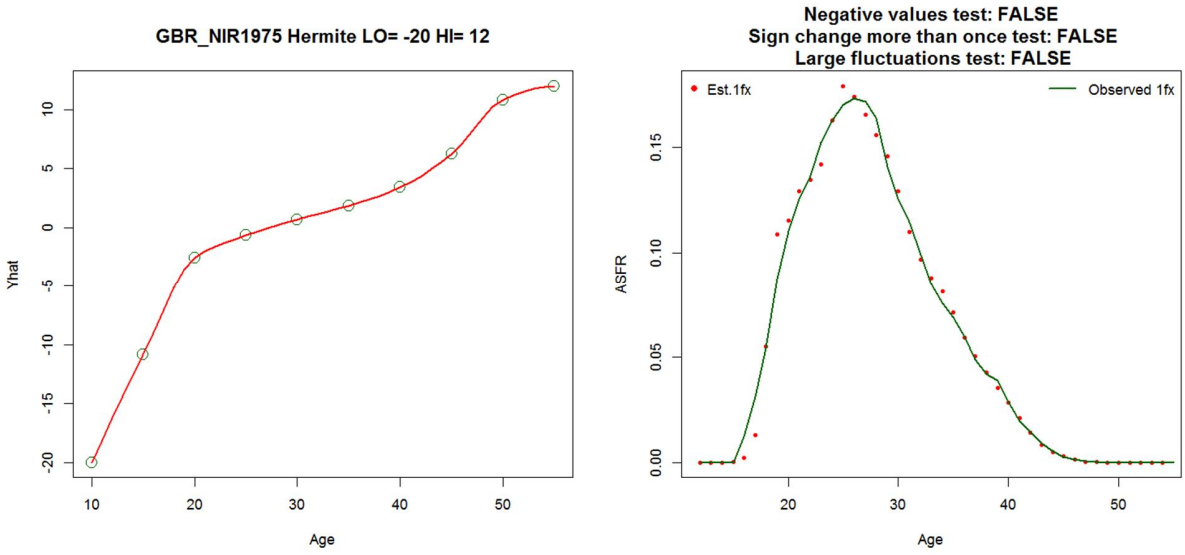


Figure 2.3
Estimating single-year age-specific fertility rates ($_1f_x$) on the basis of 5-year ($_5f_x$) age-specific fertility rates; the HFD method after adjustments, Northern Ireland, 1975

Source: As for Figure 2.1

Note that the odd patterns depicted in Figure 2.2 occurred in a case in which no fertility was recorded for the first age group. As a result, the two first values of the logit cumulative fertility curve are undefined, and have to be replaced. Setting a value of –20 is obviously too low in such cases, as it forces the cumulative curve to go too abruptly to the next value (which is much higher than –20). Assigning a small artificial value of 0.00001 to the first age group improves the results, regardless of the default values of LO and HI (Panel B of Figure 2.3).

The same experiments (adding phantom births) were repeated for other problematic cases that were generated from the real HFD data. As in the case shown above, the results improved significantly. Moreover, in each case it was possible to find empirically optimal LO and HI values that ensured that the estimated fertility curve looked plausible and close to the original single-year ASFRs. Thus, it appears that the interpolation results depend on the HFD splitting algorithm more than on the mathematical properties of the Hermite function itself, particularly in selecting optimal LO and HI values.

2.2. Calibrated spline (CS) method

The approach offered by Carl Schmertmann (2012) is quite different from the one currently used in the HFD[3] and from other disaggregation methods. The innovative component of the CS method is that the optimization task uses empirical information, which improves the plausibility of the estimated fertility curves. The objective of the CS method is to find a compromise between ***Fit*** (proximity of the predictors to the observed values) and ***Shape*** (similarity of the known fertility patterns). This goal is achieved by minimizing a squared error penalty based on these two criteria. The CS approach assumes that the optimal schedule f^* is a linear function of the observed data y. Matrix K containing predetermined constants links f^* and y:

$$f^* = K \times y$$

The core of the method is the estimation of matrix K, which involves rather complex matrix algebra. However, once K is defined, the application of the method is straightforward.

[3] The detailed documentation of the CS method, the input data, and R scripts are available online at: http://calibrated-spline.schmert.net

Our thorough testing based on the HFD and HFC data has confirmed that the CS method performs well, as measured by the smoothness of the fertility curves and the absence of sudden kinks. We also tested the CS method on birth order-specific fertility data. For these data, the *Ks* first needed to be derived. The obtained results again favored the CS estimator: even for higher birth orders, the estimated fertility schedules still looked smooth and plausible. Yet like any other method, the calibrated spline estimator has its limitations, which are primarily related to its potential usage in the HFD:

1) By construction, the CS method does not assume that the spline function should pass through defined knots. Thus, by definition it does not meet our *balance* criterion. This limitation is crucial for the HFD, in which the birth counts within aggregated age groups before and after splitting should match.
2) Occasionally, the CS method produces negative values, which mostly occur at the tails. These values are then being replaced by zero. The negative values might appear at age 15, and even at ages 16, 42, 43, or 44. The ASFRs at these ages are not very high, but are still substantial. Thus, the simple replacement of these negative values with zeroes (as is now done) results in the loss of birth counts. This method sometimes produces positive values that should be zeroes. Both problems need to be accounted for, which can be done using a simple post-correction procedure. Yet the need for such a procedure lends additional complexity to the practical implementation of this method.
3) The application of the CS method is fairly easy, but only if the input data are supplied in a uniform format. Otherwise, matrix *K* needs to be defined each time through a rather complex derivation process. It would appear that the raw data in the HFD are not standardized. The number of age groups might vary from country to country and from year to year. That implies that the matrix of constants *K* has to be estimated for each particular case.
4) When applying the CS method to the birth order data, order-specific *Ks* have to be derived. Again, this adds complexity. More importantly, birth order fertility data are scarcer, which might be an issue when constructing an empirical basis for *K*, particularly for higher birth orders.

Given these limitations, we did not consider using the CS method in the HFD. However, the CS method has proven to be a very good fit for the HFC, which contains very heterogeneous and noisy data.

3. Quadratic optimization (QO) method

Here, we present a new approach that is very different from both the HFD and CS methods. Let us consider the problem of estimating the age-specific fertility rates by single year of age from data collected by five-year age groups of the mother's age. When we analyze the number of births by five-year age groups, we are solving an empirical approximation for an integral equation, which links the age-specific fertility rates to the number of births by five-year age groups. Denote $E(x)$ number of women of age *x* in the given year in the population of interest, and $fr(x)$ – the age-specific fertility rate. If the age is continuous, the number of births $b(y)$ to women not older than *y* years is given by expression:

$$b(y) = \int_{y_0}^{y} E(x)\, fr(x)\, dx \tag{3.1}$$

where y_0 is the age when fertility starts. From (3.1) follows that the number of births b_i in age group $y_{i-1} \leq y < y_i$ for $i = 1,\ldots,n_y$ equals:

$$b_i = \int_{y_{i-1}}^{y_i} E(x)\, fr(x)\, dx \tag{3.2}$$

The objective of the QO method is to solve a quadratic optimization problem with a set of constrains regarding forms of inequalities and equalities. The method for solving the set of equations (3.2) in different settings is described in the respective sections. Section 3.1 describes the implementation of the QO method in the case when the age-specific fertility rates by single year of age are estimated from the number of births aggregated in groups. Section 3.2 describes the algorithm of estimating the age-specific fertility rates from the aggregated fertility rates. Finally, section 3.3 deals with estimating conditional and unconditional fertility rates by parity.

3.1. Estimating the age-specific fertility rates from the aggregated number of births

Approximate the set of equations (3.2) via substitution function $E(x)$ by vector E, with elements E_i being equal to the number of women in the *i-th* one-year age group, and with function $fr(x)$

by vector fr with elements fr_i being equal to the one-year age-specific fertility rate. Integration in (3.2) is replaced by a summation to obtain a matrix equation:

$$b = G' \; fr \tag{3.3}$$

where b is a vector with elements b_i equal to the number of births in the *i-th* age group, and G is an aggregator matrix with elements

$$G_{ij} = \begin{cases} E_j & j \in D_i \\ 0 & otherwise \end{cases},$$

Here, D_i is a set of years that belong to the i-th interval of the women's age. In the data presented in the HFD, there are 9 aggregated and 43 one-year age groups covering the age interval 12 to 55. This implies that matrix G is composed of 9 rows and 43 columns.

From a theoretical point of view, matrix equation (3.3) has an infinite number of solutions. In practical applications, it is a common approach to take one of the possible solutions that has a positive property, such as the smooth solution (Williams, 2013). This can be done by minimizing the quadratic functional

$$\left\| b - G' \; fr \right\|^2 + l \left\| F' \; fr \right\|^2 \to \min_{fr} \tag{3.4}$$

where $\left\| x \right\|^2 = \sum_{i=1}^{n} x_i^2$, n - number of elements in vector x, $l > 0$, F - matrix with elements

$$F_{ji} = \begin{cases} 1 & i = j, i = j+2 \\ -2 & i = j+1 \\ 0 & otherwise \end{cases}, \; i = 1, \ldots, m-2; \; j = 1, \ldots, m.$$

Here, m is the number of elements in vector fr equal to the number of years for which the age-specific fertility rate is calculated. The multiplication of vector fr by this matrix gives the vector of the second set of differences for vector fr, while the value $\left\| F' \; fr \right\|^2$ reflects the 'smoothness' of the age-specific fertility rate estimate. To get a non-negative fertility rate estimate, we should minimize (3.4) under constraints $fr_j \geq 0$, $j = 1, \ldots, m$.

In general, parameter l in (3.4) scales the 'smoothness' of the age-specific fertility rate estimate and the precision of the solution to matrix equation (3.4). If we are certain that the registered number of births b_i, i=1,..., n_y is determined with negligible error, then the first term in (3.4) can be omitted, and the new task of quadratic minimization can be written in the following form:

$$\text{minimize} \quad \|F \cdot fr\|^2$$

$$\text{subject to} \quad G \cdot fr - b = 0 \qquad (3.5)$$

$$fr_j \geq 0,\ j = 1, \ldots, m .$$

If there are no births in the first ($j=1$) and the last ($j=m$) age groups, then the first and the last columns in matrices G and F are to be omitted.

3.2. Estimating the age-specific fertility rates from the aggregated fertility rates

Fertility rate Fr_i obtained from aggregated data for age group $y_{i-1} \leq y < y_i$ is calculated by formula $Fr_i = u_i / \int_{y_{i-1}}^{y_i} E(x)dx$; here, u_i is the number of births from the i-th age group, and $\int_{y_{i-1}}^{y_i} E(x)dx$ is the total number of women. Keeping equation (3.2) in mind, we can write:

$$Fr_i \cdot \int_{y_{i-1}}^{y_i} E(x)dx = \int_{y_{i-1}}^{y_i} E(x)\, fr(x)dx .$$

If function $E(x)$ changes slowly within age interval $y_{i-1} \leq y < y_i$, the approximation for this equation can be derived in the following form:

$$Fr_i = \frac{1}{y_i - y_{i-1}} \int_{y_{i-1}}^{y_i} fr(x)dx .$$

The estimation of the age-specific fertility rate from aggregated data can be performed using the method described in section 3.1, with small modifications:

$$\text{minimize} \quad \|F \cdot fr\|^2$$

subject to $G^{*\prime}\, fr - FR = 0$

$$fr_j \geq 0, \quad j = 1, \ldots, m$$

Here, *FR* is a vector of aggregated fertility rates. The elements of matrix *G** are calculated as:

$$G^*_{ij} = \begin{cases} 1/|D_i| & j \in D_i \\ 0 & otherwise \end{cases},$$

where $|D_i|$ is the number of years in the *i*-th interval of age.

3.3. Estimating the age-specific fertility rates by parity

3.3.1. Maintaining the balance by parity within aggregated age groups

The age-specific fertility rate by parity satisfies a set of equations similar to equation (3.2)

$$b_i = \int_{y_{i-1}}^{y_i} E(x)\, fr(x)\, dx$$

$$b_i^p = \int_{y_{i-1}}^{y_i} E^p(x)\, fr^p(x)\, dx, \quad p = 1, \ldots, n_p \tag{3.6}$$

Here, *p* is the birth order (parity), $E^p(x)$ is the number of women at age *x* who had already gave *p*-1 births , $fr^p(x)$ is the respective age-specific fertility rate, and n_p is the maximal number of births. Each equation from (3.6) can be solved using the QO method described in section 3.1 by substituting the proper b_i^p and $E^p(x)$, as well as the b_i and *E(x).*

Parity-specific age distributions $E^p(x)$ are hard to obtain in practice, but they can be substituted by a distribution for all women (regardless their parity): $\tilde{E}(x) = \sum_{p=1}^{n_p} E^p(x)$. The solution of the new set of equations

$$b_i^p = \int_{y_{i-1}}^{y_i} \tilde{E}(x)\, \tilde{fr}^p(x)\, dx, \quad p = 1, \ldots, n_p \tag{3.7}$$

gives unconditional estimates for the age-specific fertility rate for a given birth order, which can be found by applying the QO method described in section 3.1. The quadratic minimization problem for solving equation (3.7) for different *p* takes the following form:

minimize $\left\| F' \; \tilde{fr}^{p} \right\|^2$

subject to $G' \; \tilde{fr}^{p} - b^{p} = 0$

$$\tilde{fr}_i^{p} \geq 0, \; i = 1, \ldots, m.$$

3.3.2. Maintaining the balance by parity within one-year age groups

Equations (3.6) and (3.7) assume that the projected numbers of births in aggregated age groups are equal to the given numbers of births in aggregated age groups b_i^p. This approach guarantees that the sum of the projected numbers of births in aggregated age groups with parities 1,...,n_p equals to the total projected numbers of births in aggregated age groups. However, this approach does not guarantee that the sum of the projected numbers of births in one-year age groups with parities 1,...,n_p is equal to the total projected numbers of births in one-year age groups. To achieve such balance, some modifications of the quadratic optimization method should be made. The quadratic minimization problem for estimation in this case is as follows:

minimize $\left\| F' \; fr \right\|^2 + \sum_{p=1}^{n_p} \left\| F' \; fr^{p} \right\|^2$

subject to $G' \; fr - b = 0$

$$G^{p\prime} \; fr^{p} - b^{p} = 0 \qquad p = 1, \ldots, n_p$$

$$fr_i \geq 0 \qquad i = 1, \ldots, m$$

$$fr_i^{p} \geq 0 \qquad i = 1, \ldots, m; \; p = 1, \ldots, n_p$$

$$E_i \, fr_i - \sum_{p=1}^{n_p} E_i^{p\prime} \; fr_i^{p} = 0 \qquad i = 1, \ldots, m$$

The last equation corresponds to the one-year age group's balance condition.

Here, elements of matrix G^{p} are calculated by the formula:

$$G_{ij}^{p} = \begin{cases} E_j^p & j \in D_i \\ 0 & otherwise \end{cases}.$$

If there are unconditional estimates for the age-specific fertility rate, all matrixes G^p are equal to matrix *G*.

4. Neural network (NN) method

Below we provide a very brief and general description of the neural network (NN) algorithm. A more detailed description of this process can be found elsewhere (Riedmiller and Braun, 1994). Our base model is a pre-learned neural network with a sigmoidal activation function evaluated with a resilient back propagation algorithm. The outputs of the network are smoothed estimates adjusted for negative values.

The neural network method is a mathematical model that builds on an analogy to the neural networks in the human brain. A biological neuron is a cell that accepts input signals, recalculates, and then sends output to other neurons. It is a complex system, the mathematical model of which has yet to be fully studied. The most important model that mathematically describes the biological neuron is a formal neuron. Networks built with formal neurons can approximate the multidimensional function on the output.

4.1. Neural networks and resilient back-propagation algorithms

The most basic part of a neural net is a formal neuron, shown in Figure 4.1.

Figure 4.1. Formal neuron

Source: Adopted from Zaencev (1999)

A formal neuron contains a weighted sum and a non-linear element (activation function). The operation of a formal neuron can be described by the following formulas:

$$NET = \sum_i w_i x_i \qquad (4.1)$$

$$OUT = F(NET - \theta), \tag{4.2}$$

where

x_i – input signals, the vector of all input values stands for x;

w_i – weight coefficients;

NET – the weighted sum of all input values, *NET* value is transferred to non-linear elements;

q – the threshold level of the neuron;

F – the non-linear function stands for the activation function – the sigmoidal (logistic) function that accepts and recalculates $(NET - \theta)$

$OUT = \frac{1}{1 + e^{-NET}}$ - Output with a sigmoidal activation function; θ equals zero for this function

The combination of formal neurons builds a neural net, shown in Figure 4.2.

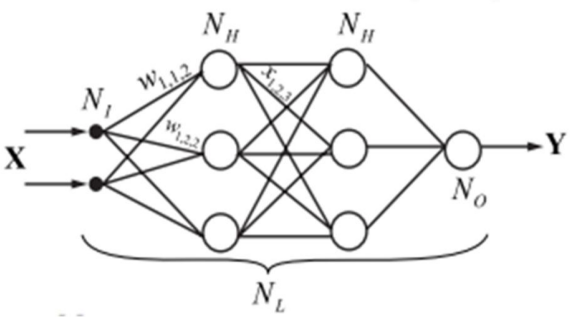


Figure 4.2. Formal structure of a neural network

Source: Adopted from Zaencev (1999)

Formally, the neural network is just a sequential evaluation of linear and non-linear function combinations:

$$f(x) = F\left(\sum_{i_k} w_{i_N j_N N} \ldots \sum_{i_2} w_{i_2 j_2 2} F\left(\sum_{i_k} w i_{1 j_1 1} - q_{j_1 1}\right) - q_{j_2 2} \ldots - q_{j_K N}\right) \tag{4.3}$$

The sequential evaluation provides a close approximation of the multidimensional function.

The resilient back-propagation algorithm (Riedmiller&Braun,1993) is a method used to tune the weights w_i in such a way that the function $f(x)$ could approximate data. After all of the train data are given to the neural net and all of the derivative errors of the formal neurons are counted, update values $\Delta^{(t)}_{ij}$ for the neural net coefficients are calculated by the system of the following equations:

$$\Delta^{(t)}_{ij} = \begin{cases} \eta^{+} * \Delta^{(t-1)}_{ij} & , \quad if \ \frac{\partial E^{(t-1)}}{\partial w_{ij}} * \frac{\partial E^{(t)}}{\partial w_{ij}} > 0 \\ \eta^{-} * \Delta^{(t-1)}_{ij} & , \quad if \ \frac{\partial E^{(t-1)}}{\partial w_{ij}} * \frac{\partial E^{(t)}}{\partial w_{ij}} < 0 \\ \Delta^{(t-1)}_{ij} & , \quad else \end{cases} \tag{4.4}$$

where

η is a step parameter. The updated values of weights are based upon the information on a local error derivative

$$\Delta w^{(t)}_{ij} = \begin{cases} -\Delta^{(t)}_{ij} & , \quad if \ \frac{\partial E^{(t)}}{\partial w_{ij}} > 0 \\ +\Delta^{(t)}_{ij} & , \quad if \ \frac{\partial E^{(t)}}{\partial w_{ij}} < 0 \\ 0 & , \quad else \end{cases} \tag{4.5}$$

Weights are updated by the following rule:

$$w^{(t+1)}_{ij} = w^{(t)}_{ij} + \Delta w^{(t)}_{ij} \tag{4.6}$$

The cycle continues until convergence is reached.

4.2. Spline and the elimination of negative values procedures

The smoothed values $\hat{f}$ of the neural net estimates f is a minimizer for the functional:

$$\sum_{i=1}^{n} (Y_i - \hat{f}(x_i))^2 + \lambda \int_{x_1}^{x_n} \hat{f}''(x)^2 dx \to \min_{\hat{f}}, \tag{4.7}$$

where $\lambda \geq 0$ is a smoothing parameter that trades the quality of the approximation and the smoothness of function (4.3).

The vector of smoothed spline parameters is calculated by the following formula:

$$\hat{m} = (I + \lambda A)^{-1} Y \tag{4.8}$$

with matrix

$$A = \int f_i''(x)\ f_j''(x)\, dx \tag{4.9}$$

where $f_i(x), f_j(x)$ are a set of spline basis functions.

The negative values are eliminated by the simple substitution:

$$\hat{f}(x_i) = \begin{cases} \hat{f}(x_i) & ,if\ \hat{f}(x_i) \geq 0 \\ 0 & if\ \hat{f}(x_i) < 0 \end{cases} \tag{4.10}$$

4.3. Disaggregation of the neural network structure

We constructed a net with nine input neurons (age groups: 10-14, 15-19, … , 50-54), two hidden layers with 27 and 36 neurons, and 43 output neurons (single age groups: 12,13,14,…,54). The structure of this neural net is given in Figure 4.3.

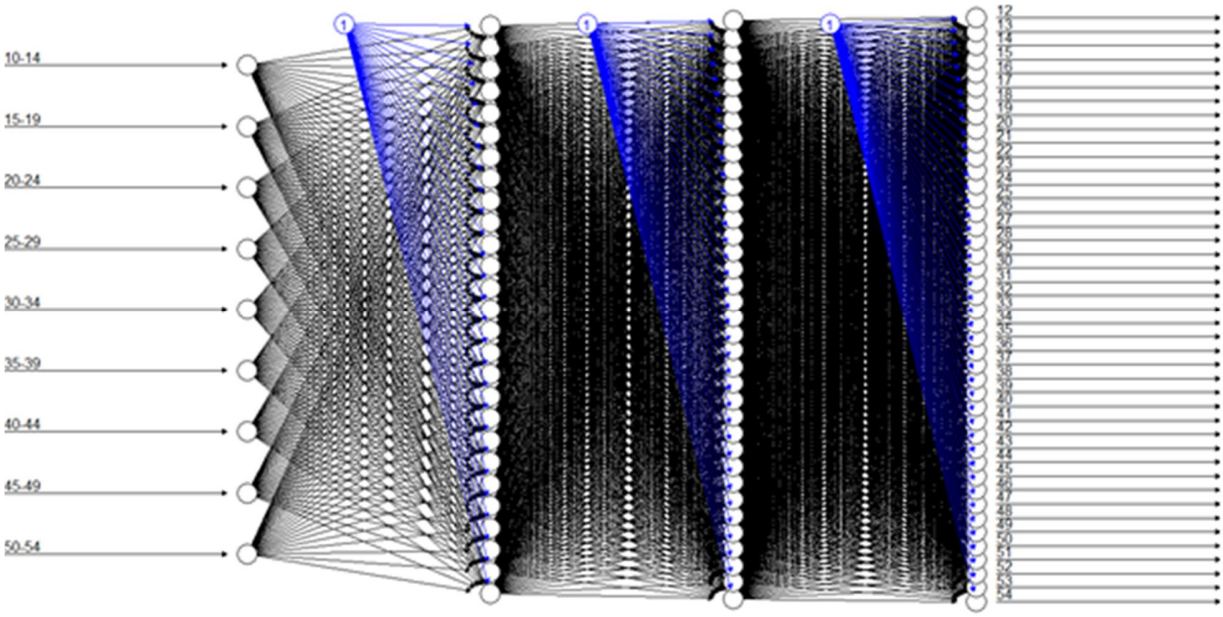


Figure 4.3. Example structure of the neural net with three hidden layers

5. Results of testing[4]

The accuracy levels of the QO and NN methods were evaluated both statistically and visually by examining the obtained fertility schedules. This was done using both the HFD and HFC data[5]. Our new methods were also tested against the current HFD splitting protocol and the CS method. This section provides a short summary of this testing procedure. The full results are presented in the supplementary materials for this paper. For the comparative analysis, we used original input data from the HFD by single year of age (birth counts and population exposures) to calculate five-year age-specific fertility rates (ASFR). These data were then split into one-year age groups using different disaggregation methods, and compared with the original ASFRs. Nine age groups (10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, and 50-54) and 43 single age groups (from 12 to 54) were used for our testing procedure. In total, we had access to 1,968 fertility schedules from the HFD. We used only the data for which original birth counts and population exposures were available by single years of age. A random sample (50 percent) was drawn from this dataset to generate the pre-learned model for the NN method. Other 50 percent samples (984 schedules) were then used in the comparative analysis of the

[4] R scripts containing QO and NN functions, fertility data, and examples are provided in the respective MPIDR Technical Reports (see Michalski, Grigoriev, Gorlischev, 2018 and Gorlischev, Grigoriev, Michalski, 2018)
[5] The results of testing using HFC data are not shown here, but are provided in Gorlischev, Grigoriev, Michalski (2018)

performance of the four methods: QO, NN, HFD, and CS. We used RMSE (root mean squared error between observed and predicted ASFRs) as a statistical indicator of model fit. For testing the performance of the QO method by parity, we relied exclusively on the visual examination of the predicted and observed fertility schedules (see supplementary file S2). We previously verified that the QO algorithm returned fertility rates that were 'balanced' both within five-year age groups and by parity.

5.1. General assessment

Figure 5.1 depicts the RMSE by country-year (984 HFD schedules) produced by different disaggregation methods. Overall, the error is below 0.01 for all HFD countries and methods. There are, however, a few exceptions, such as two cases for the HFD for which the RMSE is above 0.03 (ISL 1965 and ISL 1963). The top five outliers for each of the methods are listed in the top-right corner of the panels of Figure 5.1. Iceland 1965 (1963) appears to be the 'worst' case in terms of *fit* for all of the methods except the QO method.

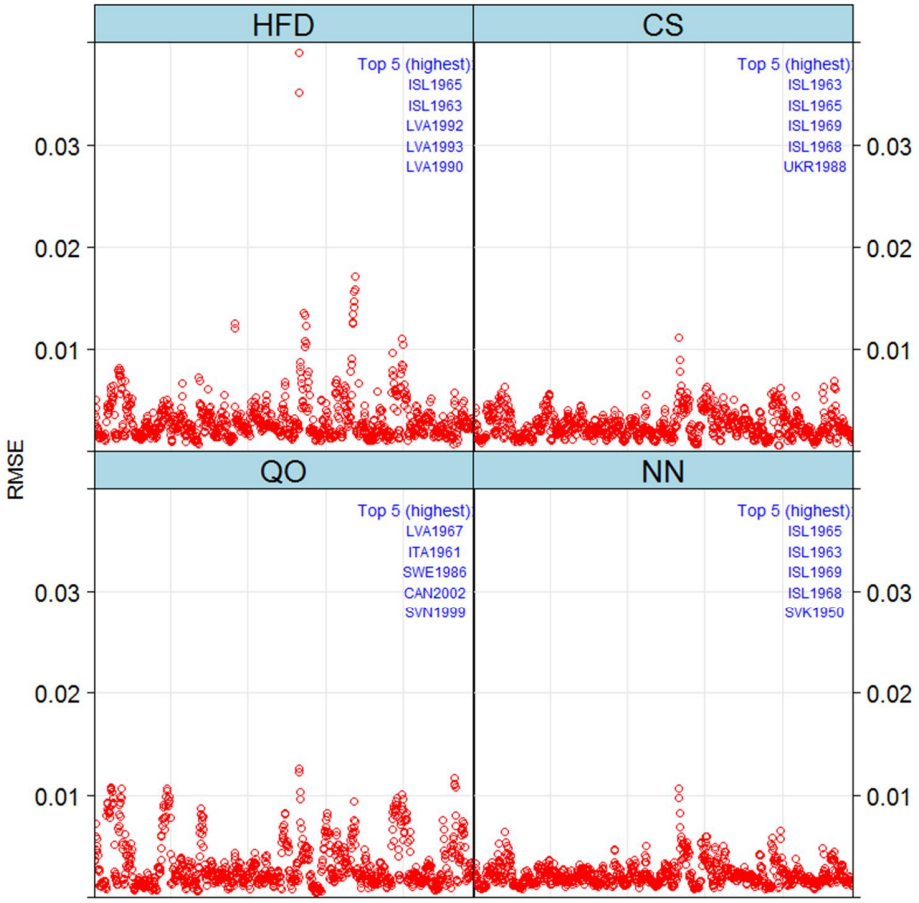


Figure 5.1. RMSE between predicted and observed age-specific fertility rates by country-year; HFD, CS, QO, and NN methods[6]

Source: Own calculations based on the HFD data
Note: country abbreviations as in the HFD

The distribution of the RMSE by age appears in Figure 5.2. Both the HFD and the QO methods return the biggest errors (particularly at ages 19 and 20), which is the 'price' for compiling an important constraint. Both methods fulfill the *balance* criterion, whereas the CS and NN methods do not. Figure 5.3 depicts the cumulative RMSE by age. We can see that the error rises very rapidly at younger ages for the HFD and QO methods; while for the CS and NN methods, the patterns appear to be smoother. There is almost no difference between the HFD and QO methods in terms of *fit* up to age 30. At higher ages, the HFD method appears to perform better. However, by construction, the QO method returns a much smoother *shape*. Out of the four methods, the NN method appears to deliver the best performance in terms of *fit*.

[6] For individual values of RMSE, see supplementary file S8

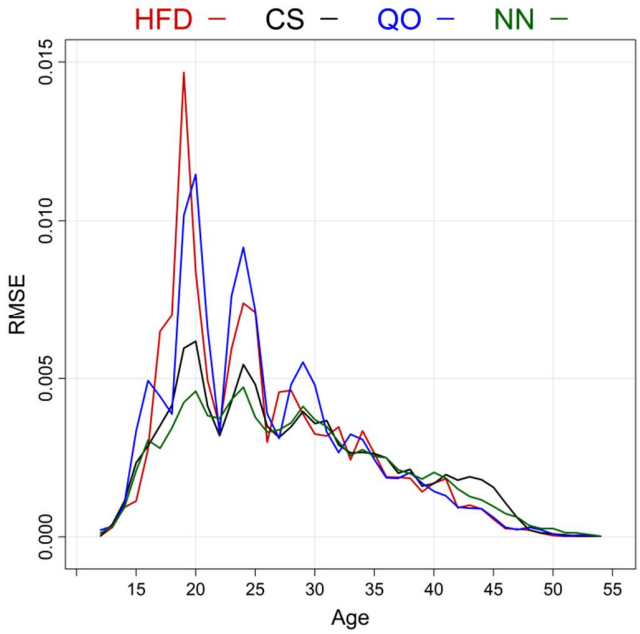


Figure 5.2. RMSE between predicted and observed age-specific fertility rates by age; HFD, CS, QO, and NN methods

Source: As for Figure 5.1
Note: average of 984 fertility schedules from the HFD

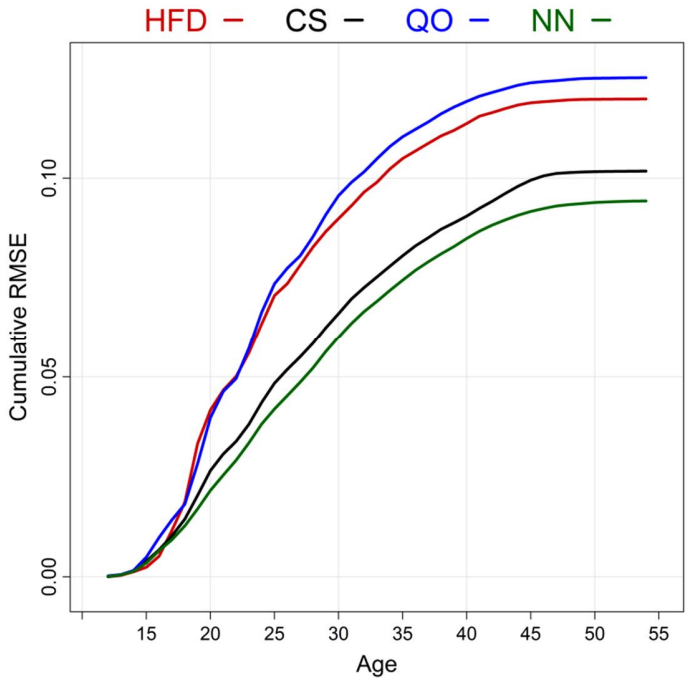


Figure 5.3. Cumulative RMSE between the predicted and the observed age-specific fertility rates by age; HFD, CS, QO, and NN methods

Source: As for Figure 5.1
Note: average of 984 fertility schedules from the HFD

5.2. Specific cases

Figure 5.4 visualizes the observed and the predicted values obtained using different disaggregation methods for several hypothetical cases[7].

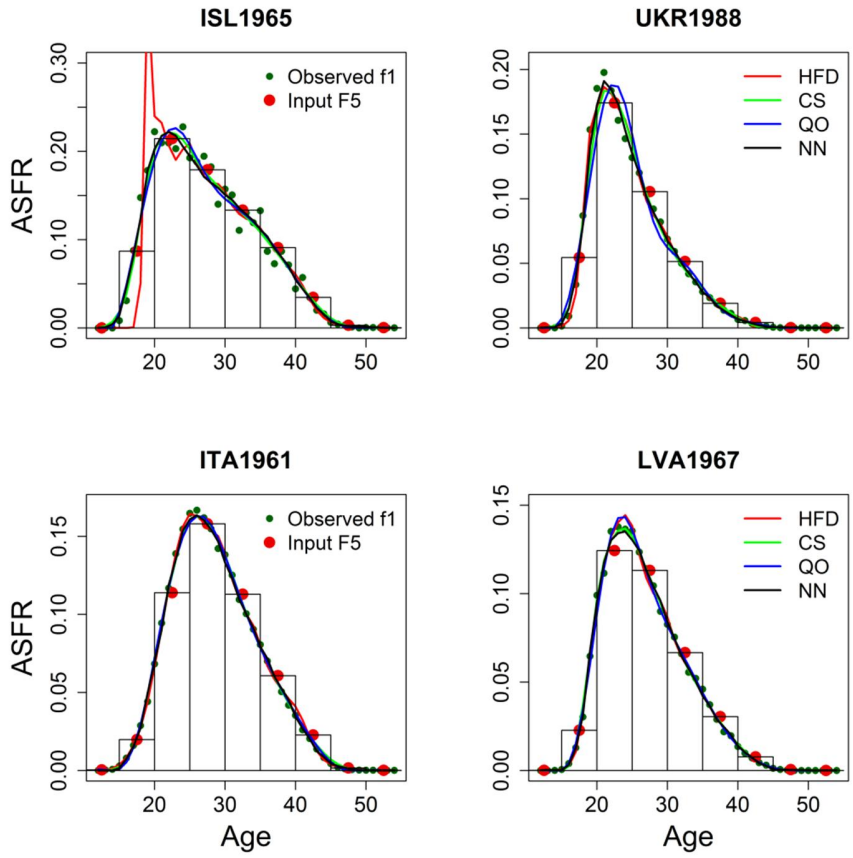


Figure 5.4. Observed and predicted values of ASFRs obtained using different disaggregation methods; selected country-years[8]

Source: As for Figure 5.1

For this example, we selected the cases with the highest RMSE (see Figure 5.1). The case of Iceland is very specific. Because the country has a small population and a small number of births, Iceland's fertility rates are very unstable. Moreover, there are no births in the first age interval. As we have already shown for the case of Northern Ireland 1975 (Figure 2.2), the HFD method is particularly unreliable in such situations. The QO and NN methods (and the CS method) do not have a similar problem: regardless of the input data, they return plausible shapes of fertility curves.

[7] These cases are hypothetical, as in the HFD the raw data for these country-years are available by single year of age
[8] See supplementary file S1 for all results

For the other cases presented in Figure 5.4 (Ukraine 1988, Italy 1988, and Latvia 1967), all of the methods look reasonable, despite the high RMSE. This finding suggests that the RMSE alone cannot be used as the main performance criterion. It should be complemented by the visual examination of the predicted and observed values. Supplementary file S1 provides such a diagnostic tool for all of the HFD country-years used in our testing.

Figure 5.5 shows an example (Austria, 2004) of estimates for unconditional age-specific fertility rates by parity relative to observed values:

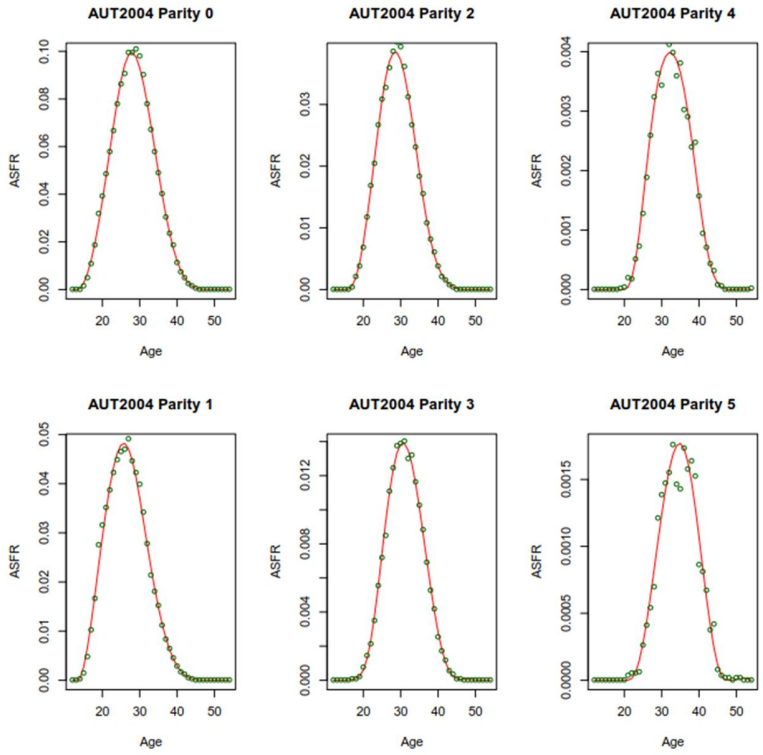


Figure 5.5. Age-specific fertility rates by parity estimated by the QO method (red line) relative to the original (green dots) values; Austria 2004

Source: As for Figure 5.1

The thorough examination of the HFD fertility schedules for which original data by birth order (parity) were available by single year of age (N=1146) suggests that the QO method performs well (see supplementary file S2). It maintains the balance with five-year age groups and by parity in most of the cases. Also, the QO method returns plausible fertility curves. However, it might sometimes produce implausible shapes for parity 1 (e.g., BGR1986, supplementary file S2). For such cases, we have developed specific solutions that can easily be adapted for other

cases. At a practical level, the method has been implemented in the form of built-in function *AGW* of R function *QOSplitPar.R* (see Michalski, Grigoriev, Gorlischev, 2018).

6. Conclusion

In this report, we considered two alternatives to the existing disaggregation methods: namely, the quadratic optimization (QO) method and the neural network (NN) method. Both methods produce reasonably accurate results. The QO method produces a smooth ASFR, which guarantees that the predicted number of births by five-year age groups is equal to the observed values. Furthermore, the QO method can be also used for ASFR estimation by parity. In some cases, however, the QO method produces relatively large errors for the ages with high fertility levels. This problem can be fixed by introducing an age weighting in the quadratic function. Nevertheless, the QO approach meets all the criteria of the 'best' splitting method, and can thus be considered a good alternative to the HFD method. While the neural network method generates highly accurate results, it requires large computational resources, and should be pre-learned. It can also give wrong estimates for types of data that were not presented in the learning data. On the other hand, the neural network approach is a relatively flexible and precise method that can be used to reconstruct missing values. Further improvements to the neural network method can be made through the optimization of some of the parameters: e.g., the number of hidden layers, the number of neurons, and the activation function for the neural net.

Supplementary materials (wp-2018-001-supplemental_materials.zip)

S1. Observed and predicted ASFRs obtained using different disaggregation methods (Observed vs predicted ASFRs by different methods.pdf)
S2. Observed (green dots) and predicted (red line) ASFRs by parity obtained using the quadratic optimization method (Observed vs predicted ASFRs by parity QO method.pdf)
S3. Original HFD data by single years of age (Original HFD data 1x1.csv)
S4. Original HFD data by 5-year age groups (Original HFD data 5x1.csv)
S5. Original HFD data by single years of age and parity (Original HFD data by parity 1x1.csv)
S6. Original HFD data by 5-year age groups and parity (Original HFD data by parity 5x1.csv)
S7. Root mean squared error (RMSE) between the observed and the predicted ASFRs by age; different disaggregation methods (RMSE different methods by age.csv)
S8. Root mean squared error (RMSE) between the observed and the predicted ASFRs by country-year; different disaggregation methods (RMSE different methods by country-year.csv)

Acknowledgements

The participation of Dmitri A. Jdanov and Vladimir M. Shkolnikov was partly supported by the Russian Academic Excellence Project «5-100».

References

de Beer C (2011). A new relational method for smoothing and projecting age-specific fertility rates: TOPALS. *Demographic Research* 24(18): 409-454.

Gorlischev V, Grigoriev P, Michalski AI. (2018). R programs for splitting abridged fertility data into a fine grid of ages using the neural network method. MPIDR Technical Report TR-2018-001. Rostock. Available at http://www.demogr.mpg.de/tr-2018-001

Grigoriev P, Jdanov DA (2015). Splitting abridged fertility data using different interpolation methods: is there the optimal solution? Presentation at the 80th PAA Meeting, San Diego, USA. http://www.humanfertility.org/Docs/paa/Grigoriev_Jdanov.pdf

Grigorieva O, Jasilioniene A, Jdanov DA, et al (2015). Methods Protocol for the Human Fertility Collection. http://www.fertilitydata.org/docs/methods.pdf

Human Fertility Database (HFD). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available at http://www.humanfertility.org

Human Fertility Collection (HFC). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available at http://www.fertilitydata.org

Jasilioniene A, Jdanov DA, Sobotka T, et al. (2012). Methods Protocol for the Human Fertility Database. http://www.humanfertility.org/Docs/methods.pdf

Liu Y, Gerland P, Spoorenberg T, Kantorova V, Andreev K (2011). Graduation methods to derive age-specific fertility rates from abridged data: a comparison of 10 methods using HFD data. Presentation at the First HFD Symposium, MPIDR, Rostock, Germany. http://www.humanfertility.org/Docs/Symposium/Liu-Gerland%20et%20al.pdf

McNeil, Donald R, Trussell TJ, Turner JC (1977). Spline interpolation of demographic data. *Demography* 14(2): 245–252.

Michalski AI, Grigoriev P, Gorlischev V (2018). R programs for splitting abridged fertility data into a fine grid of ages using the quadratic optimization method. MPIDR Technical Report TR-2018-002.Rostock. Available at http://www.demogr.mpg.de/tr-2018-002

Riedmiller M, Braun H (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In Proceedings of the IEEE International Conference on Neural Networks.

Schmertmann C (2012). Calibrated spline estimation of detailed fertility schedules from abridged data. *MPIDR Working Paper WP-2012-022*. Rostock, Germany.

Smith L, Hyndman R, Wood S. (2004). Spline interpolation for demographic variables: the monotonicity problem. *Journal of Population Research* 21 (1), pp. 95–97.

Williams HP (2013). Model Building in Mathematical Programming. Chichester: John Wiley & Sons.

Zaencev (1999). Neural networks: main models. The manual for the course "Neural networks" [Neironnye seti: osnovnye modeli. Uchebnoe posobie k kursu “Neironnye seti”]. Voronezh, 76 p.